The Trolley, the Bull Bar, and Why Engineers Should Care About the Ethics of Autonomous Cars

By JEAN-FRANÇOIS BONNEFON

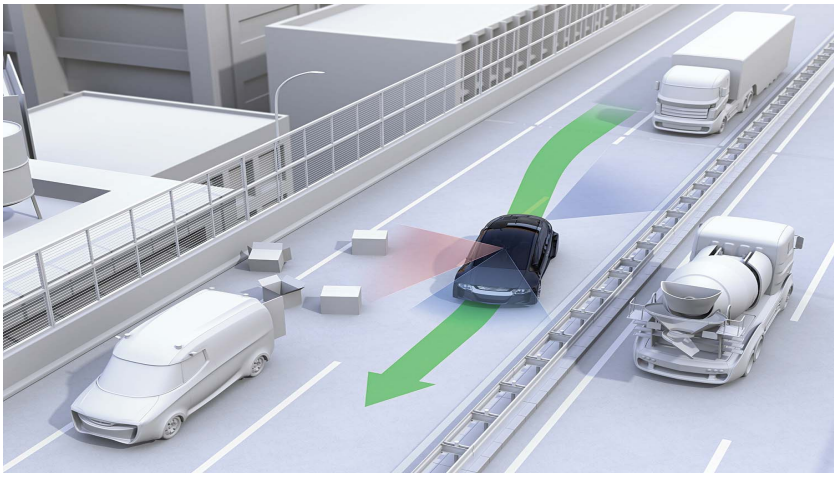
Toulouse School of Economics (TSM-R, CNRS), Université Toulouse-1 Capitole, 31000 Toulouse, France

AZIM SHARIFF

Department of Psychology, University of British Columbia, Vancouver, BC V6T1Z4, Canada

IYAD RAHWAN

The Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139 USA
Center for Humans and Machines, Max-Planck Institute for Human Development, 14195 Berlin, Germany



Everyone agrees that autonomous cars ought to save lives. Even if the cars do not live up to the most optimistic estimates of eliminating 90% of traffic fatalities [1], eliminating at least *some* traffic fatalities is one of the key promises of automated driving. Indeed, the first two principles of the *German Ethics Code for Automated and Connected Vehicles* lead with this goal as a normative imperative [2].

> *The primary purpose of partly and fully automated transport systems is to improve safety for all road users. The licensing of automated systems is not justifiable unless it promises to produce at least a diminution in harm compared with human driving [. . .].*

This view amounts to considering autonomous cars as *implicit* ethical agents, in Moor's terminology [3]. Machines are implicit ethical agents when their actions impact ethical issues (such as safety) and when these actions are constrained to avoiding unethical outcomes (such as casualties). Unlike *explicit* ethical agents, implicit ones do not learn or encode ethics explicitly—and thus, they cannot autonomously arbitrate between different kinds of harm. For example, autonomous cars as implicit ethical agents strive to avoid crashes—but when a crash is unavoidable, when all trajectories are likely to end up in casualties, implicit ethical agents find themselves dumbfounded, and unable to choose among the different ethical choices.

I. SELF-DRIVING TROLLEY PROBLEM IS LUDICROUS

Such edge cases have drawn so much attention that they

Digital Object Identifier 10.1109/JPROC.2019.2897447

threaten to cannibalize the whole field of autonomous car ethics. Almost every discussion of autonomous car ethics features some version of the so-called *trolley dilemma* [4], for example:

> *An autonomous car is barreling down on five persons, and cannot stop in time to save them. The only way to save them is to swerve and crash into an obstacle, but the passenger of the car would then die. What should the car do?*

This kind of highly stylized, black-and-white thought experiment is useful as a tool for raising public awareness of complex ethical issues. It is also useful for assessing people's moral intuitions without burdening them with an intricate mechanical description of just how such an improbable situation might arise [5], [6]. But the downside of using such stylized scenarios is that their fanciful nature and lack of realism strains the credulity of engineers, making it hard for them to care. Why is the car cruising at such unsafe speed? Why can't it just brake hard? Why is it constrained to just two trajectories? How do we know for sure that the passenger is going to die? In sum, why should we care about such vague, unrealistic, improbable use cases, when our time is best spent trying to avoid them in the first place? These are legitimate questions. From an engineering perspective, situations in which autonomous cars would act as *explicit* ethical agents (making life and death decisions based on encoded ethics) seem too thorny and far-fetched to be a priority. It is far easier to treat autonomous cars as *implicit* ethical agents, whose actions are systematically oriented toward minimizing the absolute risk of a crash.

II. STATISTICAL TROLLEY DILEMMA

Alas, ignoring the challenges of autonomous vehicles as explicit ethical agents will only postpone the problem. Even if every action of an autonomous car is oriented toward minimizing the absolute risk of a crash, each action will also shift relative risk from one road user to another [7], [8]. The cars may not be making decisions between *outright* sacrificing the lives of some to preserve those of others, but they will be making decisions about who is put at marginally more risk of being sacrificed.

For example, consider the illustration in Fig. 1. An autonomous car may position itself away from a large truck, and closer to a cyclist—marginally decreasing the absolute risk of a crash, and yet marginally increasing the relative risk incurred by the cyclist. Or it could position itself away from the bicycle, and closer to the other line of the road—marginally decreasing the absolute risk to the cyclist, and yet marginally increasing the relative risks incurred by its own passengers. These are not the dramatic, life and death decisions featured in trolley dilemmas. But once they are aggregated over millions of cars driving billions of miles, these small statistical decisions add up to life and death consequences—and prompt the same questions as the trolley dilemma did. In expectation, the trolley dilemma and the statistical trolley dilemma are equivalent.

Consider the 10 000 fatalities caused by car crashes in 2015 in urban areas of the European Union. Of these, about 61% were car passengers (drivers included), and 39% were pedestrians.[1] In other words, about three passengers died for two pedestrians. Imagine now, for the sake of argument, that the 2018 statistics would show a reversal to one passenger fatality for five pedestrian fatalities. Such a shift in relative risk would likely prompt serious investigations. Much research would be conducted to identify the cause of the shift, to assess whether it constituted an unfair development for the safety of pedestrians, and to decide which regulation would be required to correct this unfairness. In essence, we would circle back to the trolley dilemma, only in a statistical format. Instead of asking whether one passenger should die to save five pedestrians in a given crash, we would ask whether one passenger should die for five pedestrians, across all the crashes recorded in a single year.

III. BIG SHINY BULL BAR VERSUS BLACK BOX ALGORITHM

As a matter of fact, we have been there before. In 1996, a report from the U.K. Transportation Research Laboratory investigated the impact of "bull bars" on pedestrian risk [9]. Bull bars are hard metal protections fitted to the front of four-wheel vehicles, which were originally used for off-road trips in wild areas. Their usefulness in urban environment was questionable, though, and they seemed especially dangerous to pedestrians since, compared to a traditional bumper, the bars' smaller and more rigid surface area concentrated and intensified the impact on a struck victim. The report estimated that, in the year 1994, bull bars were responsible for two or three additional pedestrian fatalities in the United Kingdom. This is a very small shift—indeed, it is about the smallest shift one could imagine. Nevertheless, it led to a long series of tests and regulations which culminated in a global bull bar ban in the European Union.

What the bull bar story tells us is that physical features of the car can shift the relative risks incurred by different road users, that such a shift can be recognized as an ethical tradeoff, and that some instantiations of this tradeoff can be deemed as unacceptable. Now, a bull bar is a big shiny object in front of the car, which is easy to notice, easily understood as dangerous, and easily removed. The programming of autonomous cars, in contrast, is hidden deep under the hood. The cumulative effect of many small automated decisions, meant to minimize the absolute risk of a crash,

[1] ec.europa.eu/transport/road_safety/specialist/statistics

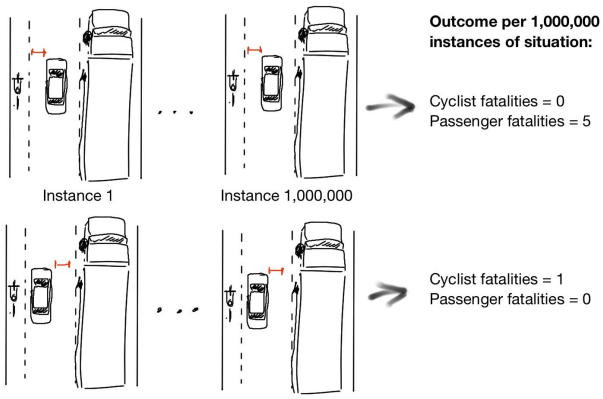


Fig. 1. ***Statistical trolley dilemma: a large truck appears in the lane next to a car with five passengers. The car must choose whether to stay put (top) or move slightly left (bottom). Aggregated over thousands of cars and millions of miles, small shifts in positioning can shift the relative risks incurred by different categories of road users (e.g., cyclists versus passengers).***

will plausibly lead to shifts in the relative risks incurred by different road users—but this cumulated effect is neither easily predictable nor easily fixable. In all likelihood, we will only learn about it after the fact, once aggregated statistics are available for all to see, assuming these statistics are up for scrutiny by independent third parties. And the question is, will we like what we see?

The problem is that we currently have no idea where to set the cursor. It is straightforward that we want autonomous cars to lead to fewer fatalities—but we do not know how these fatalities ought to be distributed among road users. We could, of course, opt for the status quo, and consider that relative risks ought to stay the same as what they are now. But there are no ethical grounds for such a choice. Current ratios of passenger to pedestrian fatalities do not result from reasoned and carefully calculated ethical forethought, but from an accumulation of split-second decisions and unexamined behaviors. The opportunity to program these decisions in advance means that we can (and thus should) make deliberate choices, but we currently have no consensus about what those choices should be.

IV. WHY THE DILEMMAS MATTER

The issue is so thorny that we might be tempted to avoid it entirely, and to consider that, well, anything goes. Let us stick to minimizing fatalities, without concerning ourselves with the distribution of these fatalities. But consider then the effect of market forces and consumer preferences. Say two competing companies market self-driving cars that both eliminate 80% of fatalities, but one company's cars split the remaining fatalities equally between passengers and pedestrians, whereas the other company's cars split the remaining fatalities nine-to-one in favor of their passengers. Consumers would flock to the cars of the second company [5], and pedestrian risks would gradually inflate to unacceptably unfair levels.

Ultimately, we will need to decide, as a society, what we consider to be a fair distribution of risks among road users. This is not an easy decision. Too much favor toward passengers over pedestrians could produce public outrage toward self-driving car companies and passengers. Too much favor toward pedestrians and potential autonomous car consumers may vote with their wallets and opt-out of yielding their autonomy to machines that do not make their owners' interests sufficiently paramount. Either scenario could derail the rollout of autonomous vehicles and delay the benefits they promise to bring. Right now, this conversation has relied heavily on trolley dilemmas, whose lack of realism has tempted many to discard the conversation as technically irrelevant. This reaction is both legitimate and mistaken. It is legitimate because trolley dilemmas do lack realism. It is mistaken because trolley dilemmas are merely the unrealistic discrete version of a very real dilemma that emerges at a statistical level. This statistical dilemma needs to be solved, and engineers must have a voice in this process. ■

REFERENCES

[1] M. Bertoncello and D. Wee, "Ten ways autonomous driving could redefine the automotive world," Inst. McKinsey Company, Tech. Rep., 2015.
[2] C. Luetge, "The German ethics code for automated and connected driving," *J. Philosophy Technol.*, vol. 30, pp. 547–558, Dec. 2017.
[3] J. H. Moor, "The nature, importance, and difficulty of machine ethics," *IEEE Intell. Syst.*, vol. 21, no. 4, pp. 18–21, Jul. 2006.
[4] P. Foot, "The problem of abortion and the doctrine of double effect," *J. Oxford Rev.*, vol. 5, pp. 5–15, 1967.
[5] J. F. Bonnefon, A. Shariff, and I. Rahwan, "The social dilemma of autonomous vehicles," *J. Sci.*, vol. 352, pp. 1573–1576, Jun. 2016.
[6] A. Shariff, J.-F. Bonnefon, and I. Rahwan, "Psychological roadblocks to the adoption of self-driving vehicles," *J. Nature Hum. Behav.*, vol. 1, pp. 694–696, Sep. 2017.
[7] N. J. Goodall, "Ethical decision making during automated vehicle crashes," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2424, no. 1, pp. 58–65, 2014.
[8] N. J. Goodall, "Away from trolley problems and toward risk management," *Appl. Artif. Intell.*, vol. 30, no. 8, pp. 810–821, 2016.
[9] B. J. Hardy, "A study of accidents involving bull bar equipped vehicles," Transp. Res. Lab., Tech. Rep. 2431996, 1996.